International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 3, Issue 5, Dec 2013, 77-84
© TJPRC Pvt. Ltd.

LIVE CAPTURING AND VIDEO TRACKING

ANURADHA BHATIA[1] & SACHIN BOJEWAR[2]

[1]PG Scholar, Department of Computer Engineering, Alamuriratnamala Institute of Engineering and Technology, Mumbai, Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering, Vidyalankar Institute of Technology, Mumbai University, Maharashtra, India

ABSTRACT

The histogram of oriented gradients descriptor is one of the best and most popular descriptorsused for pedestrian detection. Unfortunately this technique suffers from one big problem: speed. Like mostsliding window algorithms it is very slow, making it unsuitable for any realtime application.The HOG detector is a sliding window algorithm. This means that for any given imagea window is moved across at all locations and scales and a descriptor is computed. For thatwindow a pretrained classifier is used to assign a matching score to the descriptor. The classifierused is a linear SVM classifier and the descriptor is based on histograms of gradient orientations.Gradient orientations and magnitude are obtained for each pixel from the pre-processed image. If a color image is used the gradient with the maximum magnitude (and its correspondingorientation) is chosen. This is done by convoluting the image with the 1-D centred kernel[-1 0 1] by rows and columns. Several other techniques have been tried (including 3x3 Sobelmask or 2x2 diagonal masks) but the simple 1-D centred kernel gave the best performance (for pedestrian detection).The first is the idea ofusing appearance information, whose importance in object detection has been widelydemonstrated [26, 28], and specially the idea of combining cues with the HOG descriptor, e.g., co-occurrence HOG [29], color HOG [24], etc. It has been largely demonstrated by the proposal of different descriptors that appearance is an important cue forobject detection. The second source is the increasing trend of using segmentation forboth detection and segmentation, which in our case has the potential of highlighting theshape of the human.

KEYWORDS: HOG, SVM, Image Segmentation, Video Segmentation

INTRODUCTION

This project is implemented for the domain specific human detection and the image retrieval in video by exploring the techniques for feature extraction, semantic annotation and query processing. Feature extraction and semantic annotation are the main requirements to show that the indexing method is feasible; that is the information can be extracted without too much manual intervention. Query processing that includes developing a data model, choosing a suitable query language and constructing a compact representation, is the most important aspect to show the success of the indexing method in terms of satisfying users/applications' requirements.

Detecting humans in images is a challenging task owingto their variable appearance and the wide range of poses thatthey can adopt. The firrst need is a robust feature set thatallows the human form to be discriminated cleanly, even incluttered backgrounds under difficult illumination. We studythe issue of feature sets for human detection, showing that locally normalized Histogram of Oriented Gradient (HOG) descriptors provide excellent performance relative to other existing feature sets including wavelets [17, 22]. The proposed Descriptors are reminiscent of edge orientation histograms

[4, 5], SIFT descriptors [12] and shape contexts [1], but theyare computed on a dense grid of uniformly spaced cells andthey use overlapping local contrast normalizations for improved performance. We make a detailed study of the effectsof various implementation choices on detector performance, taking pedestrian detection" (the detection of mostly visiblepeople in more or less upright poses) as a test case. For simplicity and speed, we use linear SVM as a baseline classifierthroughout the study. The new detectors give essentially perfect results on the MIT pedestrian test set ,so we havecreated a more challenging set containing over 1800 pedestrian images with a large range of poses and backgrounds.On-going work suggests that our feature set performs equallywell for other shape-based object classes

HOG DESCRIPTOR AND USAGE FOR OBJECT DETECTION

The histogram of oriented algorithm for object detection was introduced in [1]. An exampleof detection results for pedestrians and heads is visible in Figure 1.

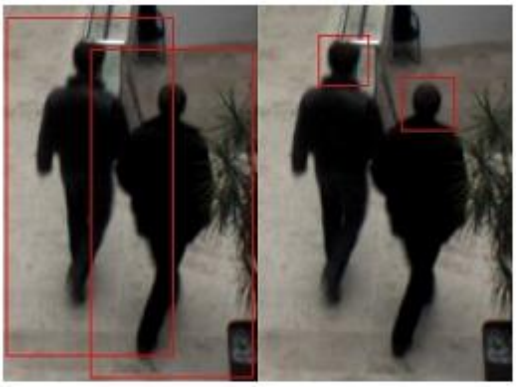

Figure 1: Example Results after Applying the HOG Detector for Pedestrians (Left) and Heads (Right)

The HOG detector is a sliding window algorithm. This means that for any given imagea window is moved across at all locations and scales and a descriptor is computed. For thatwindow a pretrained classifier is used to assign a matching score to the descriptor. The classifierused is a linear SVM classifier and the descriptor is based on histograms of gradient orientations.Firstly the image is padded and gamma normalization is applied. Padding means thatextra rows and columns of pixels are added to the image. Each new pixel gets the color ofits closest pixel from the original image. This helps the algorithm deal with the case when aperson is not fully contained inside the image. The gamma normalization has been proven toimprove performance for pedestrian detection but it may decrease performance for otherobject classes. To compute the gamma correction the color for each channel is replaced by itssquare root.

Gradient orientations and magnitude are obtained for each pixel from the pre-processed image. If a color image is used the gradient with the maximum magnitude (and its correspondingorientation) is chosen. This is done by convoluting the image with the 1-D centered kernel [-1 0 1] by rows and columns. Several other techniques have been tried (including 3x3 Sobelmask or 2x2 diagonal masks) but the simple 1-D centered kernel gave the best performance (for pedestrian detection).

Each detection window is divided into several groups of pixels, called cells. For each cell ahistogram of gradient orientations is computed. For pedestrian detection the cells are rectangular and the histogram bins are evenly spaced between 0 and 180 degrees. The contribution ofeach pixel to the cell histogram is weighted by a Gaussian centred in the middle of the blockand then interpolated trilinearly in orientation and position. This has the effect of reducingaliasing.

HOG RE-WEIGHTING USING GLOBAL IMAGE SEGMENTATION

The proposed approach consists in re-weighting the HOG descriptor for each one of thecells while it is computed. Basically, HOG consists in an intelligent grouping of gradient information (cells and blocks), as well as well-engineered histograms of gradientorientations (weighting by gradient magnitude, bin interpolation, histogram normalization and outliers clipping are the major steps1). The window of interest is covered byoverlapping blocks, therefore, the HOG descriptor of the whole window usually endsup being thousand-dimensional. A linear SVM is used for learning the human classifier works in such high dimensional space. Accordingly, the obtained classifier is just aweighted summation running on such number of dimensions.When computing HOG, the gradient orientation θPat a given pixel P is weightedby the corresponding magnitude μP, i.e., μPis accumulated in the histogram bin corresponding to θP.Notice that the gradient at P, only encodes localdifferences in intensity or color, i.e., differences between adjacent pixels. In this paper, we want to incorporate differences based on a wider spatial support into the process inorder to assess if they allow obtaining a human detector with higher performance. Inparticular, we want to weight μPby a given ωPcoming up from an image segmentation process, i.e., the vote of θPin the histogram will be the re-weighted magnitude λPinstead of μP. Figure 2 illustrates the idea.

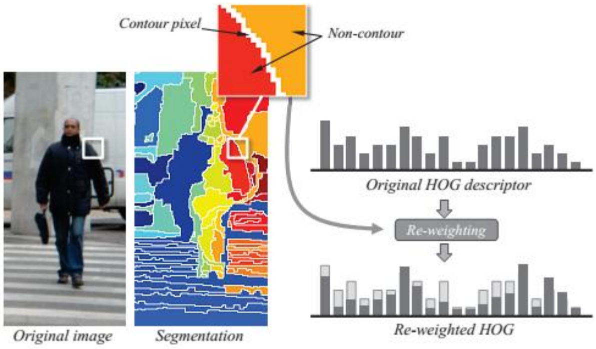

Figure 2: Re-Weighting of the HOG Descriptor According to the Image Segmentation Cues

Here, in order to compute the weighted HOG, we apply the selected image segmentation algorithm (i.e., S) to each level of the pyramid. Thus, we obtain a sort ofmulti-scale segmentation of the original image (Figure 3).

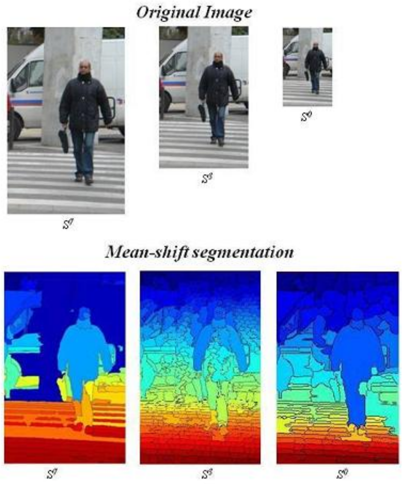

Figure 3: Pyramid-Segmentation. The Scale of the Slice in the Pyramid Affects the Segmentation, Similarly as it Affects the HOG Descriptor

PROPOSED WORK

The proposed work is divided into two parts

- The first part detects and counts the total number of visitors entering in specified domain.
- The second part of the proposed work extracts an image from an video and retrieves more videos related to the image using FP growth and SVM.

This section gives an overview of our feature extractionchain, which is summarized in Figure 1. Implementation detailsare postponed until x6. The method is based on evaluatingwell-normalized local histograms of image gradient orientations in a dense grid. Similar features have seen increasinguse over the past decade The basic idea is thatlocal object appearance and shape can often be characterizedrather well by the distribution of local intensity gradients oredge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this is implemented by dividing the image window into small spatialregions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over thepixels of the cell. The combined histogram entries form therepresentation. For better in variance to illumination, shadowing, etc., it is also useful to contrast-normalize the localresponses before using them. This can be done by accumulating a measure of local histogram energy" over somewhatlarger spatial regions ("blocks") and using the results to normalize all of the cells in the block. We will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors. Tiling the detection window witha dense (in fact, overlapping) grid of HOG descriptors andusing the combined feature vector in a conventional SVMbased window classifier gives our human detection chain in Figure 4.

The use of orientation histograms has many precursors, but it only reached maturity when combined withlocal spatial histogramming and normalization in Lowe's Scale Invariant Feature Transformation (SIFT) approach towide baseline image matching, in which it providesthe underlying image patch descriptor for matching scalein variant key-points. SIFT -style approaches perform remarkably well in this application. The Shape Contextwork studied alternative cell and block shapes, albeit initially using only edge pixel counts without the orientationhistogramming that makes the representation so effective.

The success of these sparse feature based representations has somewhat overshadowed

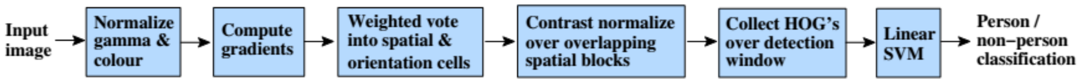
Input image
Normalize gamma & colour
Compute gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection window
Linear SVM
Person / non–person classification

Figure 4: Human Detection Chain

The Power and Simplicity of HOG'sas Dense Image Descriptors We hope that our study will help to rectify this. In particular, our informal experiments suggest that even the best current key point based approaches arelikely to have false positive rates at least 1–2 orders of magnitude higher than our dense grid approach for human detection, mainly because none of the key point detectors that weare aware of detect human body structures reliably .

The HOG/SIFT representation has several advantages. Itcaptures edge or gradient structure that is very characteristicof local shape and it does so in a local representation withan easily controllable degree of in variance to local geometricand photometric transformations: translations or rotationsmake little difference if they are much smaller that the localspatial or orientation bin size. For human detection, rathercoarse spatial sampling, fine orientation sampling and stronglocal photometric normalization turns out to be the best strategy, presumably because it permits limbs and body segmentsto change appearance and move from side to side quite a lotprovided that they maintain a roughly upright orientation.

CLUSTERING is a natural solution to abbreviate and organize the content of a video. A preview of the video contentcan simply be generated by showing a subset of clusters or therepresentative frames of each cluster. Similarly, retrieval can beperformed in an efficient way since similar shots are indexedunder the same cluster. Regardless of theseadvantages, generalsolutions for clustering video data are a hard problem. For certainapplications, motion features dominate the clustering results; forothers, visual cues such as color and texture are more important. Moreover, for certain types of applications, decoupling ofcamera and object motions ought to be done prior to clustering.

Video retrieval techniques, to date, are mostly extended directly or indirectly from image retrieval techniques. Examplesinclude first selecting key frames from shots and then extractingimage features such as color and texture features from thosekey frames for indexing and retrieval. The success from such extension, however, is doubtful since the spatial–temporal relationship among video frames is not fully exploited. Due to this consideration, recently more works have been dedicated to addressmore specifically, to exploit andutilize the motion information along the temporal dimension forretrieval.

INSERTED CAPTION DETECTION

During or after a key event in most sport videos, a text with surrounding box is usually inserted to draw the attention of users to some content-sensitive information. For example, after a goal is scored in a soccer game, some texts are usually displayed to inform viewers about the current score-line. To detect text in video, Wernicke and Lienhart used the gradient of color image to calculate the complex-values from edge orientation image. It is defined to map all edge orientation between 0 and 90 degrees, thereby distinguishing horizontal, diagonal and vertical lines. Similarly, Mita and Horilocalized character regions by extracting strong still edges and pixels with a stable intensity for two seconds. Strong edges are detected by Sobel filtering and verified to be standstill by comparing four consecutive gradient images. Thus, current text detection techniques detect the edges of the rectangle box formed by text regions in color video-frames and then check if the edge will stay for more than 2 seconds.

Based on these concepts, the essence of our text display detection method is that sport videos use only horizontal text in 99% of the cases, as can be seen in Figure 5. If we can detect strong horizontal line in a frame, we can locate the starting point of a text region. The main steps involved in detecting text display detection can be seen in Figure 5 and are explained below.

First, video track is segmented into one minute clip. Each frame for every one second gap is pre-processed to optimize performance by converting the color scheme to grayscale and the size is reduced to a smaller pre-set value. Second, Sobel Filter is performed to calculate the edge (gradient) of the current frame and Hough Transform is used on the gradient image to detect line spaces (R) between 0 - 1800. Threshold1 is applied on the R values to detect the potential candidates of strong lines which are usually formed by the box surrounding text displays. After these lines are detected, the system calculates the r (rho) and t (theta) values from the peak coordinates. It should be noted that r value indicates the location of the line in terms of the number of pixels from the center and t indicates the angle of the line. In order to verify that the detected lines are the candidates of text display, the lines are filtered out using two criteria:

- The absolute value of r is less than n % of the maximum y-axis, and
- The corresponding t is equal to 900 (horizontal). This n-value is regarded as threshold2.

The first criterion is important to ensure that the location of the lines is within the usual location of text display. The second criterion is to ensure that the line is horizontal because some strong horizontal lines can be detected from other area beside the text display, such as the boundary between field and crowd. Finally, for each of the lines detected the system check that their location (i.e. the r values) is consistent for at least m seconds. This m-value is regarded as threshold3.

The purpose of this check is to ensure that the location of the lines is consistent with the next frames as text displays always appear for at least 2 seconds to give viewers enough time to read. In fact, when a text display size is large and contains lots of information, it will be displayed even longer to give viewers enough time to read. Figure 2 illustrates how Sobel-Hough transform can be used to detect text.

Since there are not so many texts during a sport match (otherwise they can be distractive), it takes longer to check text display in all video frames. Instead, specific domain knowledge should be used to predict text appearances in sport videos which are quite typical from one match to another. Figure 5 shows an example of predicting text occurrences based on specific domain knowledge for soccer video. However, during the experiment to detect generic events, the system still check all frames since text display can detect some events which sometimes cannot be detected by whistle and excitement sounds. For example, whistle does not always exist during start of a match, but texts are usually displayed at the beginning of a match to show essential information like team formation.

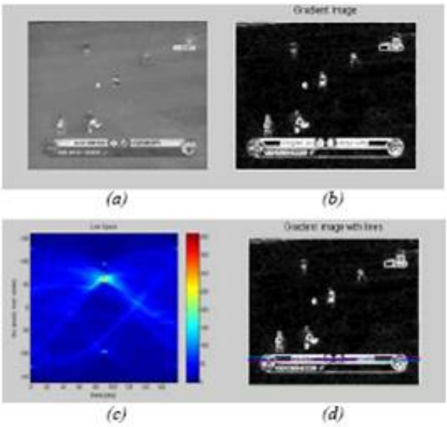

Figure 5: How Sobel/Hough Transform Detect Text: (A) Gray Scaled& Resized Frame, (B) Gradient Image Reveals the Lines (C) Peaks in Hough Transform, (D) Horizontal Lines Detected

CONCLUSIONS

In this work I have to investigate thepossibility of improving HOG descriptors inthecontext of human detection. Inparticularby weighting HOG with information comingfrom image segmentation. The detected images are all stored for mining to count the number of people in a specific domain.The specific events detection can be improved and extended by deriving some additional statistical features such as the location of slow motion replay of the last close-up frame, and including more mid-level features which can improve the highlights detection.

The work will also retrieve the videos from net on the retrieved image on the video as video tracking.

REFERENCES

1. Arbel´ aez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical imagesegmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence99(1), 898–916(2010).

2. Boix, X., Gonfaus, J., van de Weijer, J., Bagdanov, A., Serrat, J., Gonz` alez, J.: Harmonypotentials for joint classification and segmentation. In: IEEE Conf. on Computer Vision andPattern Recognition. San Francisco, CA, USA (2010).

3. Carreira, J, Sminchisescu, C.: Constrained parametric min-cuts for automatic object segmentation (2010).

4. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEETrans. On Pattern Analysis and Machine Intelligence24(5), 603–619 (2002).

5. Dalal, N.: Finding people in images and videos. PhD Thesis, Institut National Polytechniquede Grenoble / INRIA Rh ˆ one-Alpes (2006).

6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf.on Computer Vision and Pattern Recognition. San Diego, CA, USA (2005).

7. F. Bergeaud and S. Mallat. Matching pursuit of images. InImage Processing, 1995. Proceedings, International Conference on, volume 1, pages 53–56. IEEE, 1995.

8. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, andA. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303–338, June 2010.

9. Z. Rahman, D. Jobson, and G. Woodell. Multi-scale retinexfor color image enhancement. In Image Processing, 1996.Proceedings, International Conference on, volume 3, pages1003–1006. IEEE, 1996